Response Sheet

Comprehensive Inter-comparison of Generative AI Models for Super-Resolution Precipitation Downscaling Across Hydroclimatic Regimes

Shivam Singh*[1,2] Simon Michael Papalexiou[3,4,7] Hebatallah M. Abdelmoaty[3,5,] Tom Hartvigsen[6], Antonios Mamalakis[2,6]

Reviewer 1

This manuscript presents a timely and meaningful intercomparison of three "most used" gen AI models for precipitation super-resolution downscaling. The study addresses an important problem at the interface of atmospheric science and machine learning, and the effort to compare multiple model classes within a unified framework is valuable. At the same time, several aspects of the manuscript require substantial clarification and strengthening before the conclusions can be fully supported. Addressing these issues would improve the scientific rigor and impact of the paper.

Response: We sincerely thank the reviewer for the careful and insightful evaluation of our manuscript. We appreciate the recognition of the importance of the problem and the value of a unified comparison framework. We have carefully revised the manuscript to address all comments raised, including clarifications to the experimental design, additional analyses requested by the reviewer, revisions to figures and results, and expanded discussion of the study limitations and implications. Detailed responses to each comment are provided below.

Major Comments

Comment 1:

The authors state that the 10-member ensemble is generated from 10 independently trained models initialized with different random seeds. This procedure primarily reflects epistemic uncertainty associated with parameter estimation and training variability. However, the central theoretical advantage of conditional generative models is that for a given low-resolution input, they can generate a distribution of plausible high-resolution outputs through stochastic sampling. At present, the manuscript uses one prediction from each independently trained model and interprets the resulting spread as ensemble uncertainty, which is not equivalent to sampling the conditional output distribution of a single trained generative model. The authors must separate these two uncertainty sources explicitly. In addition to the current analysis, they should report results from repeated stochastic sampling using a single trained model, preferably the best-performing checkpoint, and compare that spread with the spread arising from different training seeds. This distinction is essential for a correct interpretation of the ensemble results.

Response: We thank the reviewer for this important observation. We agree that uncertainty arising from stochastic sampling of a single trained generative model (aleatoric variability) is conceptually distinct from uncertainty arising from independently trained model initializations (epistemic variability). To address this concern, we have added a new subsection entitled **"5.2.6 Stochastic Variability and Uncertainty"** in the revised manuscript (Lines 692–726). In this analysis, we explicitly compare (i) within-seed variability obtained from repeated stochastic realizations generated by a single trained model and (ii) across-seed variability obtained from independently trained models initialized with different random seeds. The comparison is performed using spatial autocorrelation, exceedance probability, and power spectral density diagnostics. The results show that both sources of variability affect precipitation statistics, but across-seed variability produces a broader and more persistent spread, particularly in the upper tail of the precipitation distribution. These findings clarify the distinct contributions of sampling variability and training variability and provide a more rigorous interpretation of ensemble uncertainty in generative precipitation downscaling models.

The updated text between lines 689 and 725 in the revised version is,

5.2.6 Stochastic Variability and Uncertainty

The comparison between within-seed and across-seed ensembles shows that both sources of variability affect the simulated precipitation distribution, but across-seed variability is consistently larger. This distinction is most evident in the exceedance probability curves (Figure 9b1–b2). Within-seed realizations remain relatively clustered at moderate thresholds but begin to diverge in the upper tail, particularly beyond approximately 150 mm/day. In contrast, across-seed ensembles show a broader spread even at lower precipitation thresholds, and this spread increases further toward the most extreme events. This indicates that stochastic sampling within a fixed trained model contributes to tail uncertainty, but differences among independently trained model initializations produce a stronger and more persistent variability. The lagged spatial autocorrelation characteristics were highly consistent across independent model initializations (Supplementary Figure S8), indicating that the learned spatial dependence structure is robust to training variability. Although the spread among realizations increased slightly at larger lags, particularly for DDPM, the median correlations remained stable across seeds and closely followed the target spatial autocorrelation decay.

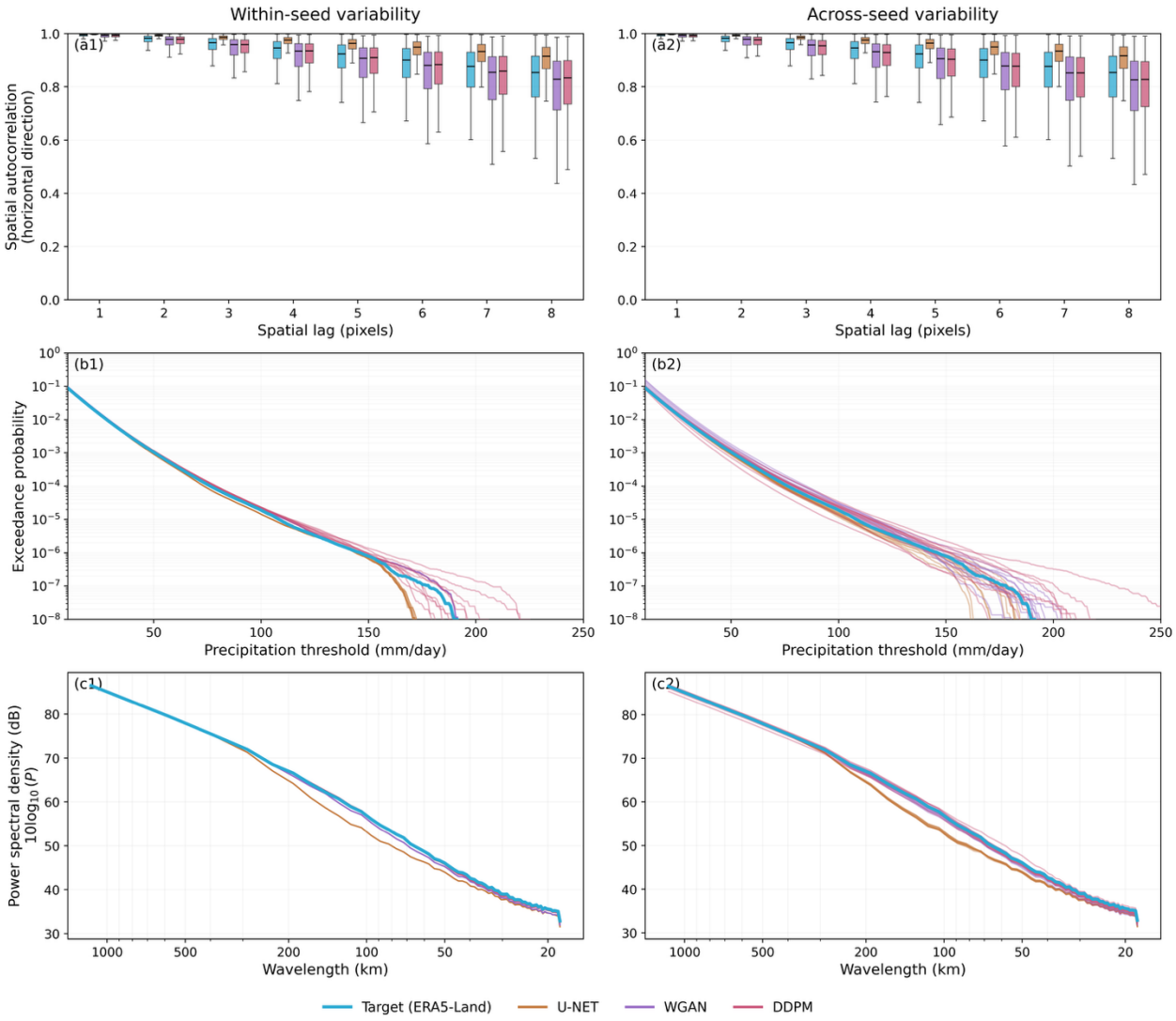

Figure 9: Within-seed and across-seed variability of U-Net, WGAN, and DDPM predictions over the independent Northeast test region at 16× super-resolution. Panels (a1–a2) show boxplots of horizontal spatial autocorrelation at lags 1–8 pixels, where within-seed variability is estimated from multiple stochastic realizations generated by a single trained model and across-seed variability is estimated from independently trained model initializations. Panels (b1–b2) present exceedance probability curves as a function of precipitation threshold (mm/day), illustrating variability in the representation of moderate-to-extreme precipitation intensities. Panels (c1–c2) show radial power spectra describing the distribution of spatial variance across wavelengths. Colored lines and boxplots represent individual model realizations, while the target (ERA5-Land) reference is shown for comparison. Left-column panels correspond to within-seed variability and right-column panels correspond to across-seed variability.

For spatial autocorrelation (Figure 9a1–a2) and radial power spectra (Figure 9c1–c2), the contrast between within-seed and across-seed behavior is less pronounced. Both ensemble configurations show broadly similar spatial dependence and spectral

characteristics, suggesting that the learned spatial organization is relatively stable. Nevertheless, the across-seed boxplots show slightly larger dispersion at longer spatial lags, indicating that model initialization also affects spatial dependence, although less strongly than it affects precipitation extremes. These results suggest that uncertainty in generated precipitation fields is metric dependent. Spatial structure is comparatively robust to both stochastic sampling and independent model initialization, whereas distributional tail behavior is more sensitive, with across-seed variability representing the dominant source of uncertainty."

Comment 2:

The manuscript refers in several places to ERA5-Land as "observation". This terminology is not correct. ERA5-Land is a reanalysis-based product, not a direct observational dataset. Since the study does not use in situ station obs, radar, satellite retrievals, or soundings as reference truth, the manuscript should consistently refer to ERA5-Land as a reanalysis or reanalysis-based target, not as observation. You could read this paper to obtain the detailed reason.

https://doi.org/10.1175/BAMS-D-14-00226.1

Response: We thank the reviewer for pointing out this terminology issue. We agree that ERA5-Land is a reanalysis product and not a direct observational dataset. In the original manuscript, the term "observed" was used to denote the reference target against which model outputs were evaluated. To improve accuracy and consistency, we have revised the terminology throughout the manuscript and figures, replacing references such as "Observed" with "Target (ERA5-Land)" where appropriate.

Comment 3:

The use of min-max normalization may help stabilize training, but it raises an important concern for precipitation, especially for extreme events. Min-max scaling bounds the normalized target by the range seen in the training data, which may hinder robust extrapolation to unprecedented values. This issue is especially relevant for climate-related downscaling and extreme precipitation, where out-of-sample events may exceed the historical training maximum. This issue may also be relevant to the behavior shown in Figure 2, where DDPM with T=100 approaches the upper bound and cannot grow freely. The authors should discuss this limitation explicitly and test at least one alternative normalization strategy such as quantile normalization or z-score normalization over wet pixels only, reporting whether the normalization choice materially changes the extreme-value results.

Response: We thank the reviewer for raising this important point. We agree that normalization can influence the representation of extremes and that the implications of min-max scaling deserve careful consideration in climate downscaling applications. To evaluate the sensitivity of our conclusions to the normalization strategy, we performed an additional experiment in which the DDPM model was retrained using z-score normalization and

compared against the original min-max normalized implementation. We evaluated both models using the same diagnostics employed throughout the manuscript, including spatial autocorrelation, exceedance probability, and radial power spectrum analyses (Figure X1).

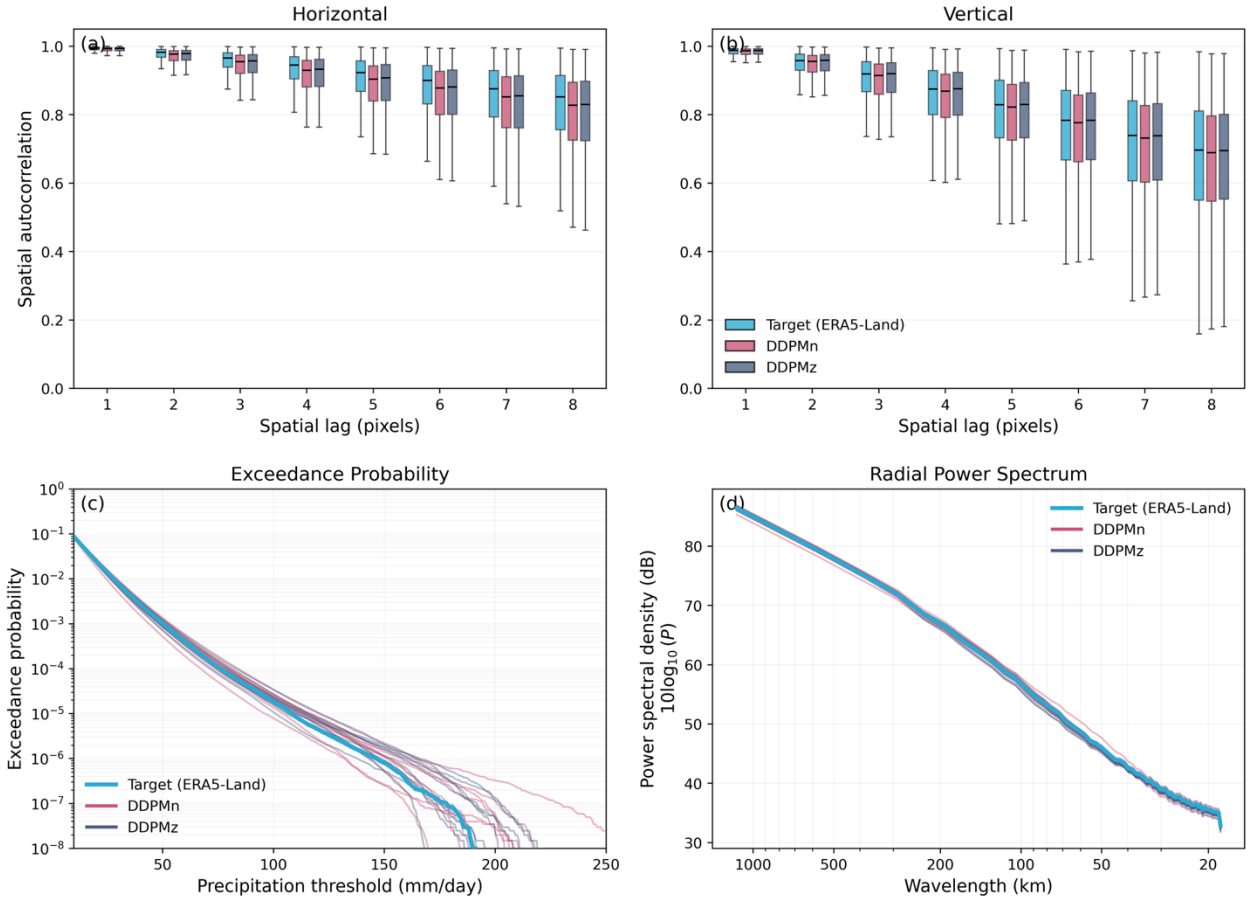

Figure X1: Comparison of spatial dependence and distributional characteristics between DDPM models trained using min-max normalization (DDPMn) and z-score normalization (DDPMz). (a) Horizontal and (b) vertical lagged spatial autocorrelation shown as boxplots across evaluation samples for lags 1–8 grid cells. Target (ERA5-Land) fields are shown in cyan, DDPMn in pink, and DDPMz in dark blue. Both models reproduce the gradual decay of spatial correlation with increasing lag and exhibit very similar autocorrelation distributions. (c) Exceedance probability curves showing the probability of precipitation intensity exceeding a given threshold. Thin colored lines represent individual DDPM ensemble members, while the cyan line denotes the target dataset. (d) Radial power spectra plotted as a function of spatial wavelength, illustrating the distribution of variance across spatial scales.

The results showed highly consistent behavior between the two normalization strategies, with only modest quantitative differences. Spatial autocorrelation distributions were nearly identical for both DDPMn and DDPMz across horizontal and vertical directions, indicating that the representation of spatial dependence is largely unaffected by the normalization choice. Similarly, exceedance probability curves from individual ensemble members showed substantial overlap between the two configurations, with z-score normalization exhibiting

marginally closer agreement to the target distribution at some thresholds. The radial power spectra were also remarkably similar across all spatial wavelengths, demonstrating that both normalization approaches reproduce comparable spatial variance distributions.

Importantly, the exceedance probability analysis did not reveal substantial differences in upper-tail behavior between the min-max and z-score normalized models within the range of events represented in the evaluation dataset. These additional experiments suggest that the principal conclusions regarding DDPM performance, spatial structure, and distributional characteristics are robust to the normalization strategy employed. While z-score normalization yielded slight improvements in some diagnostics, these differences were small relative to ensemble variability and did not materially alter the conclusions drawn from the original min-max-normalized model used throughout the study. We acknowledge this sensitivity (due to data normalization choices) to be taken into account as a future scope of the study in our conclusion section.

Comment 4:

Precipitation is not a typical continuous variable like temperature, pressure, or geopotential height. It is sparse, intermittent, highly skewed, and often better represented by zero-inflated or Tweedie-like distributions. For this reason, the loss-function choice deserves much more discussion than it currently receives. The authors should discuss why their selected losses are appropriate for precipitation specifically, and whether distribution-aware losses could improve tail behavior and wet-day occurrence. Recent studies suggest that distributional losses can be beneficial for precipitation prediction and downscaling. At minimum, this should be discussed more clearly. Ideally, the authors would include a sensitivity test or ablation experiment.

https://arxiv.org/html/2509.08369

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2024GL111828

Response: We thank the reviewer for this insightful comment. We agree that precipitation exhibits characteristics such as non-negativity, intermittency, strong skewness, and heavy-tailed behavior, which are not fully aligned with the assumptions implicit in conventional MSE-based optimization. The primary objective of the present study was to provide a controlled comparison of representative generative downscaling frameworks rather than to optimize each model using precipitation-specific loss functions. For this reason, the U-Net baseline was trained using a standard MSE objective, while the WGAN and DDPM models employed their respective conventional training objectives.

We agree that alternative loss formulations, including distribution-aware objectives such as Tweedie, Gamma, quantile-based, or other probabilistic losses, may provide advantages for representing precipitation occurrence and extremes. We have therefore expanded the discussion of study limitations and future research directions to acknowledge this issue and highlight the investigation of precipitation-specific loss functions as an important avenue for future work.

We have added this discussion in our conclusion section between line 863 and 871, "The architectures evaluated here should be viewed as representative implementations rather than fully optimized realizations. In particular, the baseline U-Net employed a conventional MSE objective, whereas precipitation fields exhibit intermittency, skewness, and heavy-tailed behavior that may be better represented by distribution-aware loss functions (e.g., Tweedie, quantile-based, or probabilistic objectives; (Hunt, 2025; Rastogi et al., 2025)). Future work should investigate such alternatives alongside improved architectures, data transformations, conditioning strategies, and physically informed constraints. Finally, all experiments were conducted under historical climate conditions, and the robustness of these approaches under future climate states and associated distributional shifts remains an important area for future investigation."

Comment 5:

Figure 5 appears to compare model predictions against an upsampled low-resolution field rather than the native high-resolution ERA5-Land target, given the clustering of identical reference values on the x-axis. If so, the comparison is not appropriate and the figure needs to be redone using the actual high-resolution target field. If this interpretation is incorrect, the authors should clarify exactly how the reference field in Figure 5 was constructed.

Response: We thank the reviewer for this observation. The clustering pattern arises because the statistics shown in Figure 5 are computed by comparing multiple model realizations against the same ERA5-Land reference samples. Consequently, identical target values appear repeatedly along the x-axis while corresponding model predictions vary slightly across realizations, producing visible clustering in the scatter plots. The figure therefore does not indicate that a lower-resolution field was used as the reference target.

In addition, following further evaluation of precipitation distribution characteristics, we revised the figure to replace conventional skewness and kurtosis with L-moment-based statistics (L-skewness and L-kurtosis). Unlike conventional higher-order moments, L-moments are generally more robust for highly skewed and heavy-tailed hydrological variables such as precipitation and are less sensitive to extreme outliers and sampling variability. This revision provides a more stable and interpretable assessment of distributional characteristics across the different model classes.

To avoid ambiguity, we have revised the figure caption and accompanying text to clarify both the use of the high-resolution ERA5-Land target fields as the reference dataset and the rationale for adopting L-moment-based diagnostics.

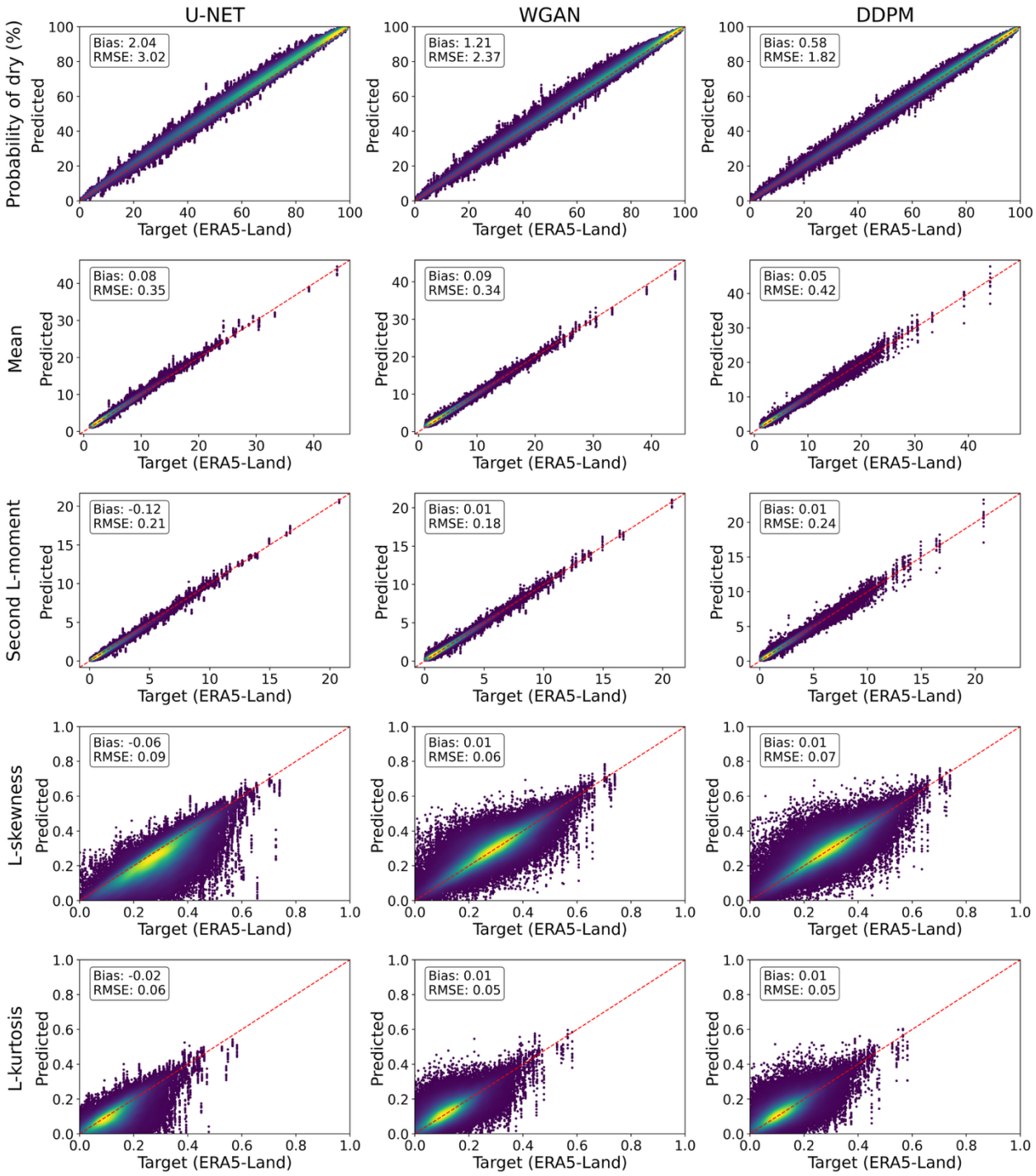

Figure 5. Comparison of precipitation statistics from high-resolution ERA5-Land reanalysis and model predictions for the 16× downscaling experiment across U-NET, WGAN, and DDPM. Each panel compares reference statistics computed from native high-resolution ERA5-Land fields (x-axis) with corresponding model predictions (y-axis), including dry-pixel probability (P_0), mean, and second to fourth central moments. Multiple seed-based predictions are evaluated against the same reference samples, leading to repeated x-axis values. Bias and RMSE are shown in each panel."

Comment 6:

The spatial lag analysis in Fig. 6 is not the most informative way to evaluate scale-dependent structure for precipitation super-resolution. A spatial power spectrum would be more standard and more physically interpretable. The authors should add a spectral analysis to the main paper. The current spatial lag figure could be moved to the supplement.

Response: We thank the reviewer for this suggestion. We agree that spectral diagnostics provide an important assessment of scale-dependent spatial structure. As noted by the reviewer, a radial power spectrum analysis was already included in the original manuscript (Figure 8a where spectral power is evaluated as a function of spatial wavelength. This diagnostic demonstrates that all models reproduce the large-scale energy distribution reasonably well while differences become more apparent at smaller spatial scales.

At the same time, we believe that the lagged spatial autocorrelation analysis provides complementary information that is not fully captured by the power spectrum alone. While the radial power spectrum characterizes the distribution of variance across spatial scales, the lagged autocorrelation analysis directly quantifies the decay of spatial dependence with distance and provides an intuitive measure of local spatial coherence. Together, these diagnostics offer a more complete assessment of spatial structure preservation. Therefore, we have retained Figure 6 (lagged spatial autocorrelation) in the main manuscript.

Comment 7:

All three models condition only on coarse-resolution precipitation. Precipitation is not a self-contained variable. For instance, topography is a key control on high-resolution precipitation structure, especially in regions where orographic effects are important. The manuscript does not sufficiently discuss the implications of omitting terrain height or other static geographic information as conditioning variables. The smoothness seen in the deterministic baseline may partly reflect the lack of physically informative conditioning, rather than only the architecture itself. This point is also relevant for the generative models. One of the strengths of conditional DDPM is the flexibility with which conditioning information can be incorporated, including modulation-based conditioning (like FiLM used here and AdaGN etc.). The authors should discuss more directly whether including terrain or other physically meaningful covariates could materially change the conclusions.

Response: We thank the reviewer for this insightful comment. We agree that physically meaningful conditioning variables, such as temperature, humidity, wind fields, topography, and land-surface characteristics, can provide valuable information for precipitation downscaling and may improve model performance. However, the primary objective of this study was to conduct a controlled intercomparison of three widely used generative AI architectures (U-NET, WGAN, and DDPM) within a perfect-model super-resolution setting. Restricting the input to coarse-resolution precipitation allowed us to isolate differences attributable to model architecture and training strategy while avoiding confounding effects arising from differing predictor selections.

We agree that incorporating physically meaningful conditioning information represents an important step toward more realistic climate downscaling applications. Such information could be incorporated through additional predictor channels or dedicated conditioning mechanisms and may further improve the representation of precipitation occurrence, intensity, and spatial structure. We have therefore expanded the discussion of study limitations and future research directions to acknowledge this point and highlight physics-informed and multivariate conditioning frameworks as promising avenues for future work.

We added the following discussion to the limitations section (Lines 858–863):

"The analysis was also restricted to precipitation, and the relative advantages of generative approaches may not necessarily extend to multivariate downscaling problems involving additional variables. Furthermore, the models did not explicitly incorporate physiographic information such as topography and land-surface characteristics, which may influence performance in regions where terrain strongly affects precipitation patterns."

We also expanded the future directions section (Lines 872–875):

"Future work should address these limitations progressively. Evaluating model performance against observation-based targets, incorporating physiographic conditioning variables, and extending the framework to multivariate downscaling problems are all important near-term directions."

Comment 8:

A plain UNet trained with a pointwise loss is well known to produce overly smooth outputs in super resolution tasks, so the trad contrast between UNet and generative models may partly reflect the baseline choice rather than an inherent limitation of deterministic approaches. The paper should justify this baseline more carefully or include at least one stronger deterministic baseline such as a sub-pixel convolution or PixelShuffle-based architecture.

https://arxiv.org/pdf/1609.05158

Response: We agree that a standard U-NET trained with a pointwise loss (e.g., MSE) is known to produce smoother outputs in super-resolution tasks and that more advanced deterministic architectures incorporating residual connections, sub-pixel convolution (PixelShuffle), attention mechanisms, or alternative upsampling strategies may improve the representation of fine-scale precipitation structure and extremes.

However, as we previously mentioned, the primary objective of this study was to conduct a controlled comparison of three widely used downscaling paradigms, deterministic regression (U-NET), adversarial generation (WGAN), and diffusion-based generation (DDPM) rather than to identify the optimal architecture within each model class. For this

reason, we selected a conventional U-NET as a simple, widely adopted, and well-established baseline. This choice allowed differences in performance to be interpreted primarily in terms of the underlying generative framework rather than architecture-specific refinements.

We nevertheless agree that stronger deterministic baselines could potentially reduce part of the performance gap observed between U-NET and the generative models. We have therefore expanded the limitations and future research directions to acknowledge that alternative architectures, improved upsampling strategies, and more advanced loss functions may further improve deterministic downscaling performance and warrant systematic investigation in future work.

We added the following discussion to the conclusion section (Lines 863–868) as limitations:

The architectures evaluated here should be viewed as representative implementations rather than fully optimized realizations. In particular, the baseline U-Net employed a conventional MSE objective, whereas precipitation fields exhibit intermittency, skewness, and heavy-tailed behavior that may be better represented by distribution-aware loss functions (e.g., Tweedie, quantile-based, or probabilistic objectives; (Hunt, 2025; Rastogi et al., 2025)).

Also see lines 875-877,

"Further research should also explore the sensitivity of conclusions to architectural design choices, loss functions, data transformations, and conditioning strategies."

Comment 9:

SSIM is a perceptual image metric designed for natural-image comparison based on luminance and contrast, and its physical meaning for sparse, intermittent precipitation fields is limited. SSIM should not be emphasized as a primary result and may be moved to the supplement. The Q-Q diagnostics currently in Figure S7 are more physically meaningful for a heavy-tailed intermittent variable and should be promoted to the main text.

Response: We thank the reviewer for this helpful suggestion. We agree that quantile–quantile (Q–Q) analysis provides a more direct assessment of precipitation distribution characteristics and extreme-value behavior than SSIM in the context of precipitation downscaling. Following the reviewer's recommendation, we have revised the manuscript by replacing the SSIM analysis in the main results with a Q–Q analysis and corresponding figure. The SSIM results have been retained in the Supplementary Material as an additional reference metric (Figure S6).

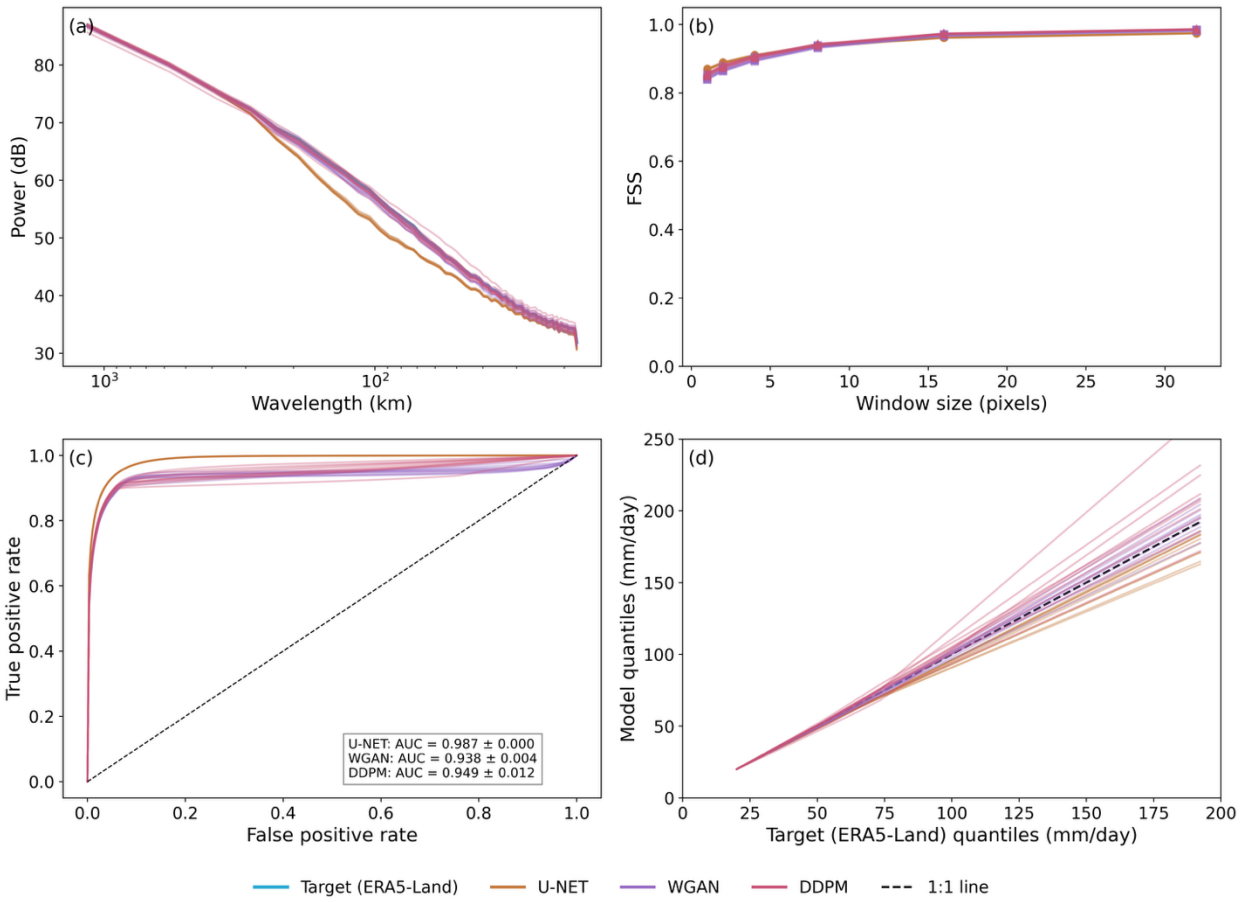

Figure 8. Comparison of U-Net, WGAN, and DDPM performance over the independent Northeast test region at 16× super-resolution. (a) Radial power spectra showing the distribution of spatial variance across wavelengths relative to the target (ERA5-Land) fields. (b) Fractions Skill Score (FSS) as a function of spatial window size. (c) Receiver Operating Characteristic (ROC) curves for wet-pixel classification, with inset values showing the mean ± standard deviation of the Area Under the Curve (AUC) across ten independent model initializations. (d) Quantile–Quantile (Q–Q) comparison between model and target precipitation quantiles for precipitation amounts exceeding 20 mm day^{-1}. Colored lines represent individual model initializations, and the dashed line in (d) denotes perfect agreement (1:1 relationship).

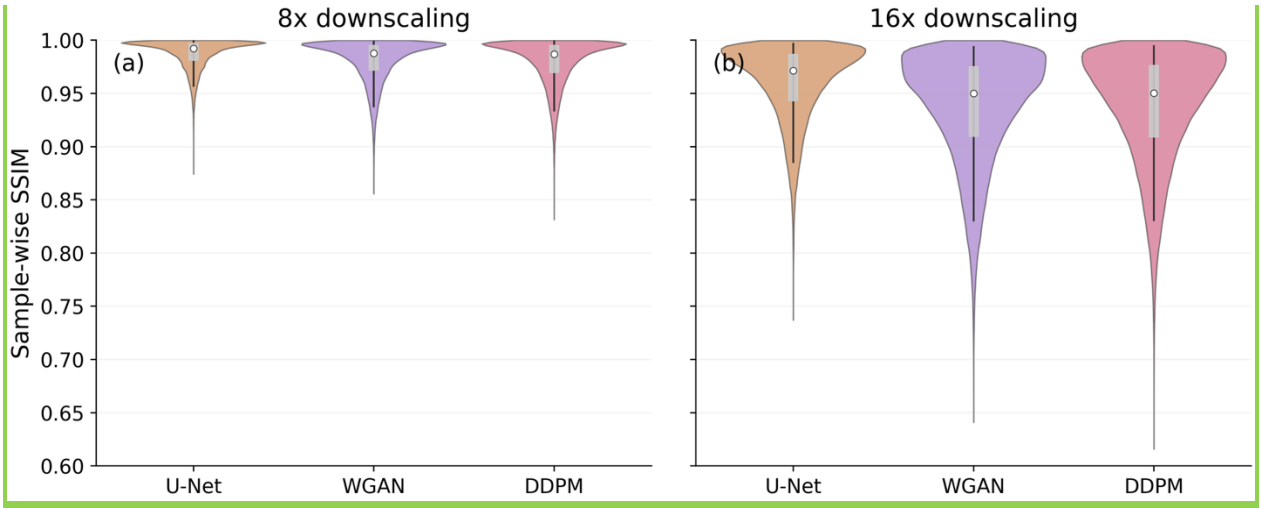

Figure S6. Distribution of sample-wise Structural Similarity Index Measure (SSIM) values for U-Net, WGAN, and DDPM across (a) 8× and (b) 16× precipitation downscaling experiments. SSIM was computed between each predicted precipitation field and the corresponding Target (ERA5-Land) field for all test samples and all 10 independently initialized seeds. Violin widths represent the density of SSIM values, while the internal vertical bar and marker denote the interquartile range (25th–75th percentile) and median, respectively. All models achieve high structural similarity at 8× downscaling, with distributions concentrated near SSIM ≈ 1.0. At 16× downscaling, the distributions broaden substantially, indicating increased difficulty in reconstructing fine-scale precipitation structures. U-Net exhibits the highest median SSIM and the narrowest distribution, whereas WGAN and DDPM display greater variability and longer lower tails, reflecting increased diversity in generated spatial patterns at higher downscaling factors.

Comment 10:

Because the low-resolution inputs are constructed using a block-averaging operator, the mass conservation result in Section 5.2.2 is partly guaranteed by the experimental design and is less informative than the manuscript implies. This section should be shortened and some of the discussion moved to the supplement.

Response: We thank the reviewer for this thoughtful comment. We agree that some degree of consistency between low-resolution and high-resolution precipitation totals is expected because the low-resolution inputs are generated through block averaging of the high-resolution reference fields. However, the model predictions are not explicitly constrained to conserve precipitation totals, and therefore deviations from the one-to-one relationship remain possible. For this reason, we view this analysis not as a strict test of mass conservation, but rather as a diagnostic of how consistently each model preserves precipitation depth across spatial scales after the super-resolution reconstruction process. To address the reviewer's concern, we have clarified the interpretation of this analysis, emphasizing that the results should be viewed as a diagnostic of consistency between

reconstructed and reference precipitation fields rather than as a strict test of mass conservation.

See lines 532-539, "The bias remains nearly constant across aggregation levels for each model. This is expected because spatial aggregation preserves the domain-averaged precipitation amount and therefore has limited influence on systematic precipitation-depth biases. Consequently, the reduction in RMSE with increasing aggregation scale primarily reflects the smoothing of local-scale errors rather than changes in the overall precipitation budget. By aggregation scales of approximately 8×8 pixels and larger, all three models exhibit near-perfect correspondence with the target precipitation depth, with correlations approaching 1.0 and only minimal residual errors."

Comment 11:

The manuscript compares a U-Net against generative models, but it does not include a simple interpolation baseline. That omission weakens the benchmark. At minimum, the authors should include one standard interpolation baseline so that readers can assess whether the deterministic neural model actually adds value beyond trivial reconstruction.

Response: We thank the reviewer for this suggestion. We agree that a simple interpolation baseline provides an important reference for assessing whether learned super-resolution models offer added value beyond straightforward spatial reconstruction. Following the reviewer's recommendation, we evaluated a bilinear interpolation baseline and included a qualitative comparison alongside the U-Net, WGAN, and DDPM results. The comparison demonstrates that bilinear interpolation produces overly smooth precipitation fields and fails to recover fine-scale spatial variability, whereas the learned models generate substantially more realistic precipitation structures. This addition helps place the model results in the context of a commonly used non-learning baseline and allows a clearer assessment of the value added by the super-resolution models.

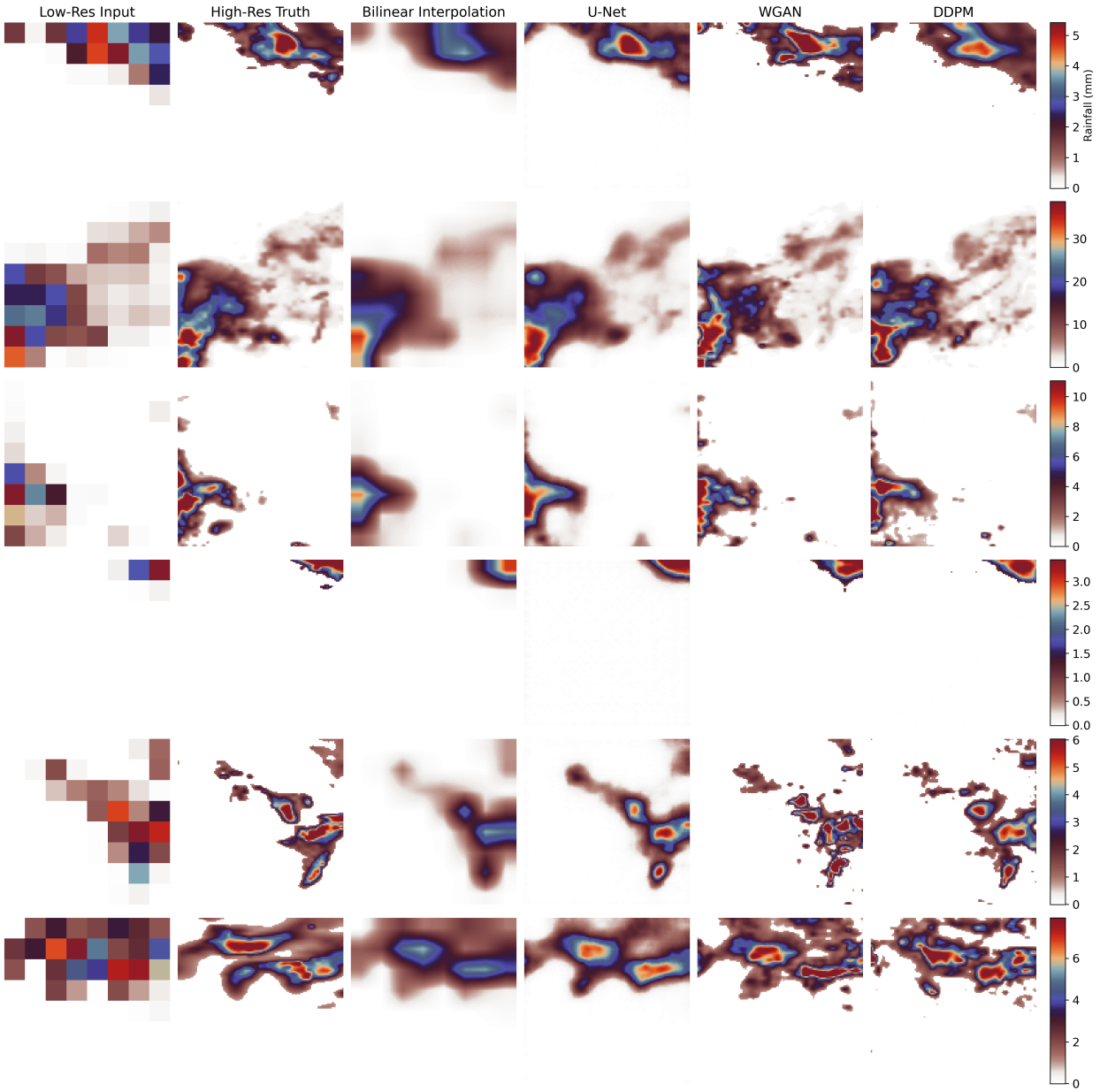

Figure S4. Representative qualitative comparison of precipitation downscaling results for five test samples from the Northeast evaluation region in the 16× super-resolution experiment. Columns show the low-resolution input (8×8), high-resolution ERA5-Land target, bilinear interpolation baseline, and predictions from the U-Net, WGAN, and DDPM models. Rows correspond to different precipitation events spanning a range of spatial structures and intensities. Bilinear interpolation produces overly smooth fields and fails to recover fine-scale precipitation features. The U-Net improves spatial detail but tends to generate smoother and more spatially coherent structures. In contrast, WGAN and DDPM recover sharper precipitation boundaries, localized high-intensity regions, and more realistic fine-scale variability, yielding spatial patterns that more closely resemble the ERA5-Land target fields.

Comment 12:

Precipitation has strong temporal autocorrelation because it is tied to evolving synoptic and mesoscale systems. Wet-spell duration, dry-spell duration, and multi-day persistence are among the most hydrologically relevant properties of any downscaled product. A model may match daily spatial structure while still failing to reproduce realistic persistence across time. The manuscript does not evaluate this sufficiently. The authors should either include temporal diagnostics that directly assess persistence behavior or clearly state that the current evaluation is not enough to establish hydrological usefulness.

Response: We thank the reviewer for this important comment. We agree that hydrological usefulness depends not only on spatial realism but also on the ability of downscaled precipitation fields to reproduce observed temporal behavior. Although the models were trained independently on daily precipitation fields without explicit temporal constraints, we conducted an additional analysis to evaluate whether the generated fields preserve temporal persistence characteristics.

To address this concern, we added a new subsection 5.2.7 (See Line 726-759),

5.2.7 Temporal Correlation

Although the downscaling models were trained independently on daily precipitation fields without explicit temporal constraints, they reproduce the observed temporal dependence structure reasonably well. As shown in Figure 10a, all three models capture the overall distribution of pixel-wise temporal correlation across lags 1–5 days, with median values closely matching those of the target (ERA5-Land). The strongest persistence is observed at lag 1, followed by a rapid reduction in autocorrelation at longer lags, consistent with the temporal behavior of the target precipitation fields.

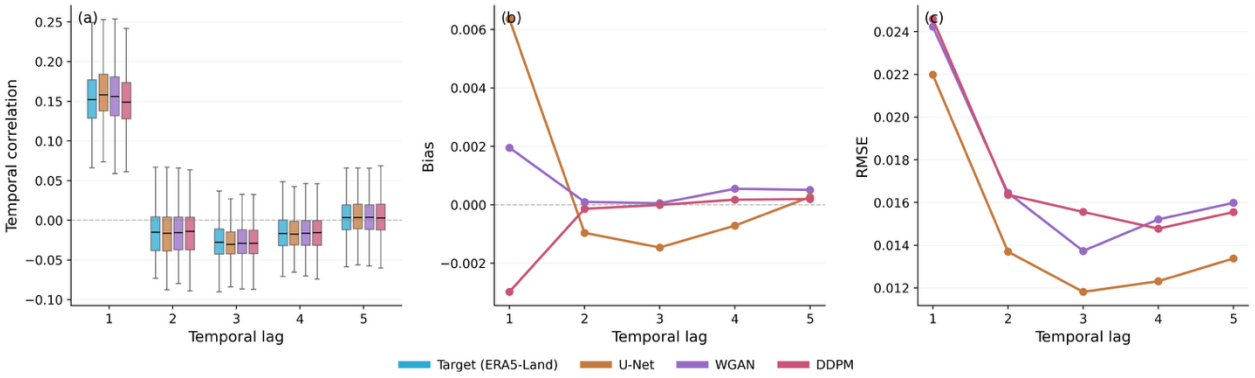

Figure 10: Temporal correlation characteristics of U-Net, WGAN, and DDPM evaluated using 1000 consecutive precipitation samples from the Northwest test region without wet-day filtering. (a) Distribution of pixel-wise temporal correlation coefficients at lags 1–5 days for the target (ERA5-Land) and model-generated precipitation fields. (b) Mean bias of temporal autocorrelation relative to the target as a function of temporal lag. (c) Root mean square error (RMSE) of temporal correlation relative to the target as a function of temporal lag.

Temporal correlation was computed independently at each grid cell using lagged correlations between daily precipitation time series.

Differences among the models become more apparent in the bias and RMSE statistics (Figure10b–c). WGAN and DDPM exhibit near-zero bias across most temporal lags, indicating that they reproduce the mean temporal correlation structure of the target data with high fidelity. In contrast, U-Net slightly overestimates persistence at lag 1 and marginally underestimates correlations at longer lags. Despite this bias pattern, U-Net achieves the lowest RMSE across all lags, indicating the closest overall agreement with the target temporal correlations. This may reflect the more deterministic and spatially smoother nature of U-Net predictions, whereas the additional variability introduced by WGAN and DDPM to better represent spatial structure and extremes results in slightly larger temporal-correlation errors while remaining nearly unbiased. Nevertheless, these differences are modest, and all three models successfully reproduce the observed decay of temporal correlation with increasing lag. The RMSE values decrease substantially from lag 1 to lag 3 before stabilizing at longer lags (Figure 10c), indicating that model differences are most pronounced in the representation of short-term persistence. These findings suggest that the spatial downscaling models preserve much of the temporal correlation structure present in the target data, despite not being explicitly trained to model temporal evolution. This provides additional evidence that the generated precipitation fields remain physically consistent not only in their spatial characteristics but also in their short-term temporal dependence.

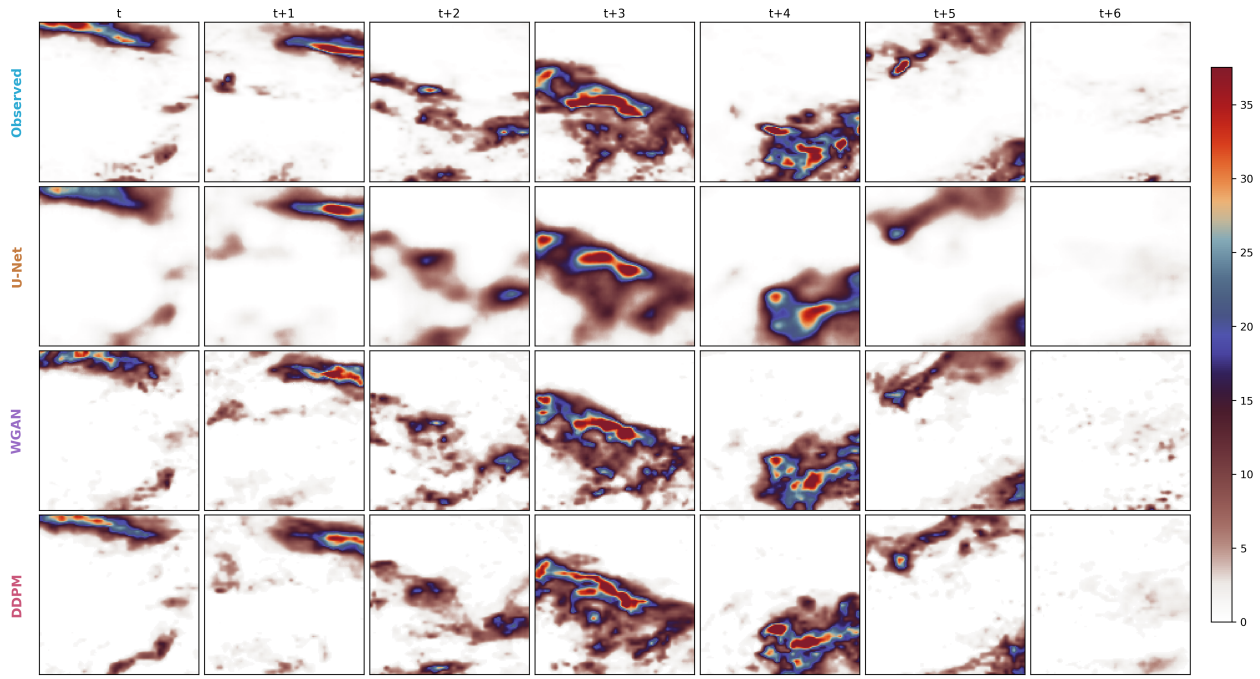

Figure X2. Reconstruction of seven randomly selected consecutive test samples by the U-Net, WGAN, and DDPM downscaling models. Columns correspond to six consecutive low-resolution input cases (t to $t + 6$), while rows show the ERA5-Land target field (Observed) and the corresponding model predictions from U-Net, WGAN, and DDPM. Color shading indicates daily rainfall intensity (mm/ day).

Minor Comments:

Comment 1:

Figures 3 and S3 should state the timestamp or time period shown. The figure captions should clearly indicate which date or sample is being plotted.

Response: Thank you for this comment. The examples shown in Figures 3 and S3 were selected randomly from the independent test set and are intended to illustrate representative reconstruction performance rather than specifically chosen events. Because the models were trained and evaluated using randomly paired low-resolution and high-resolution precipitation fields without explicit temporal information as an input or conditioning variable, the timestamp itself does not influence the model predictions.

Comment 2:

At line 126, the manuscript refers to one region as the "Pacific Northwest". Based on the domain shown, this terminology appears inaccurate, since Utah and western Nevada are not usually considered part of the Pacific Northwest. It would be better to use "Northwest" unless the domain is redefined.

Response: Thank you for your suggestion, we have replaced the region Pacific Northwest with Northwest throughout the manuscript.

Comment 3:

The manuscript states that the models are trained on the Central Plains and Northwest, while validation uses the Central Plains plus a subset of the Northeast, and the remaining Northeast samples are used for independent testing. This is understandable, but the exact fractions or sample counts should be stated explicitly in the main text.

Response: We thank the reviewer for highlighting the need for a clearer description of the dataset partitioning. Following preprocessing and quality-control filtering, 11,025 samples were retained for the Central Plains, 11,747 samples for the Northwest, and 12,348 samples for the Northeast region. In the revised manuscript, the Central Plains and Northwest datasets were used for model development (training and validation), while the entire Northeast dataset (12,348 samples) was reserved as an independent test set.

We note that an earlier version of the manuscript reported results using 12,616 test samples, which included a small number of additional samples beyond the Northeast-only evaluation

set. To improve consistency and regional independence, all figures, diagnostics, and performance metrics in the revised manuscript have been regenerated using only the Northeast test dataset (12,348 samples). This ensures that all reported results correspond exclusively to the independent Northeast evaluation region.

See lines 154-156, "After preprocessing, 11,025 samples are retained for the Central Plains, 11,747 for the Northwest, and 12,348 for the Northeast. The dataset is then split into 19,200 training samples, 3,304 validation samples, and 12,348 test samples."

Comment 4:

The manuscript sets daily precipitation below 1 mm per day to zero and excludes days with fewer than 1 percent wet pixels. These choices may be reasonable, but the authors should report how many samples are removed by region and season and briefly discuss the potential impact on light-rain statistics and wet-day occurrence.

Response: We thank the reviewer for this important comment. We agree that filtering low-precipitation samples may influence the distribution of training and evaluation data and should therefore be quantified and discussed explicitly. In the revised manuscript, we now report the fraction of samples removed by the filtering procedure for each region and season.

Overall, the filtering removes 13.76% of samples in the Central Plains, 8.11% in the Northwest, and 3.41% in the Northeast. A clear seasonal dependence is observed, with the largest removal rates occurring during winter (DJF) and autumn (SON), particularly in the Central Plains, while summer (JJA) experiences minimal filtering. These results indicate that the filtering primarily affects weak-precipitation and near-dry scenes.

We have also expanded the manuscript discussion to acknowledge that, while the filtering improves the stability of model training by focusing on meaningful precipitation structures, it may lead to an underrepresentation of light-rainfall events and a slight reduction in wet-day occurrence frequency. This potential source of bias is now explicitly discussed in the revised manuscript.

Added in main text (lines 144-153), "Daily precipitation values below 1 mm/day are treated as dry and set to zero, following commonly used wet-day thresholds in hydro-climatological analyses (Teutschbein & Seibert, 2012; Trenberth et al., 2015). To ensure that learning is driven by meaningful spatial rainfall structure rather than near-empty scenes, days with fewer than 1% wet pixels (<164 wet pixels in a 128 × 128 domain) are excluded. This filtering removes 13.76%, 8.11%, and 3.41% of samples in the Central Plains, Northwest, and Northeast regions, respectively, with the highest removal occurring during winter (DJF) and autumn (SON), and minimal removal during summer (JJA). This indicates that the filtering primarily excludes dry and weak precipitation events. While this step improves training stability, it may lead to an underrepresentation of light rainfall and a slight reduction in wet-day occurrence frequency."

Comment 5:

The manuscript states that the models are trained on the Central Plains and Northwest, while validation uses the Central Plains plus a subset of the Northeast, and the remaining Northeast samples are used for independent testing. This is understandable, but the exact fractions or sample counts should be stated explicitly in the main text.

Response: As addressed in our response to Minor Comment 3 above, "After preprocessing, 11,025 samples are retained for the Central Plains, 11,747 for the Northwest, and 12,348 for the Northeast. The dataset is then split into 19,200 training samples, 3,304 validation samples, and 12,616 test samples. (See lines 155-157 in revised manuscript)

Comment 6:

The U-Net training description omits weight decay and the Adam beta values; since the WGAN uses non-default beta1=0.0 and beta2=0.9, any deviation from defaults for other models must also be explicitly reported. The DDPM section does not specify the optimizer, initial learning rate, or weight decay.

Response: Thank you for pointing this out. We have revised the manuscript to explicitly report optimizer hyperparameters for all models. The U-Net uses Adam with default $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and no weight decay. The WGAN already used non-default values ($\beta_1 = 0.0$, $\beta_2 = 0.9$), which are now consistently reported alongside other models. The DDPM section has been updated to specify the use of AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-4}.

The text has been updated to, "The model is trained end-to-end using the Adam optimizer (learning rate 1×10^{-4}, $\text{beta}_1 = 0.9$, $\text{beta}_2 = 0.999$, batch size 32) with early stopping when validation loss fails to improve for ten consecutive epochs (patience = 10). No weight decay is used, and the best checkpoint based on validation set MSE is retained for evaluation." (See line 194-198)

Reviewer 2

I. General Comments

The study compares three deep learning frameworks (U-NET, WGAN, and DDPM) for precipitation super-resolution across different hydroclimatic regimes. While the comparison is timely, there are significant critical concerns regarding the experimental design and the technical execution that need to be addressed before the paper can be considered for publication.

Response: We sincerely thank the reviewer for the careful evaluation of our manuscript and for the constructive comments and suggestions. We appreciate the recognition of the importance of the problem while also identifying aspects of the experimental design, evaluation framework, and interpretation that required clarification. We have carefully considered all comments and substantially revised the manuscript accordingly. The revisions include clarifications to the experimental design, additional analyses and diagnostics, expanded discussion of limitations and uncertainty, and improvements to the presentation of results. Detailed responses to each comment are provided below.

II. Major Concerns

Comment 1

Training Stability and Overfitting (Critical Concern)

A fundamental issue exists regarding the training convergence and generalization of the generative models, particularly the DDPM. In Section 3 (Lines 398–400), the authors describe the dataset splitting into training and testing sets, but no validation set is mentioned. However, validation curves are provided in the Supplemental Material (Figure S2), at least for the UNET and the DDPM models. They are missing for the WGAN model. Upon inspection of Figure S2, despite the compressed scale of the y-axis, the DDPM model appears to exhibit clear signs of overfitting after approximately epoch 10. The divergence between training and validation loss suggests that the model is no longer learning generalizable features of the precipitation fields but is instead memorizing the training samples. Since the remainder of the paper relies on the results derived from these trained weights, the validity of the inter-comparison and the subsequent conclusions regarding DDPM's performance are in question. Also, the training curves of the WGAN raise some concerns, and no validation curve is shown. The authors must:

- Clarify the lack of a validation set description in the main text and for WGAN in Figure S2
- Provide a detailed analysis of the loss curves with a more appropriate y-axis scale. Is the overfitting actually happening?

- Address how they ensured the "optimal" stopping point for training to prevent reporting results from an overfit model.
- In lines 430-434, the authors acknowledge the potential for overfitting but treat it in a too simplistic manner.

Response: We thank the reviewer for this careful evaluation of the training behavior and generalization characteristics of the models. We agree that the original manuscript did not provide sufficient detail regarding the validation procedure, model selection strategy, and interpretation of the training curves.

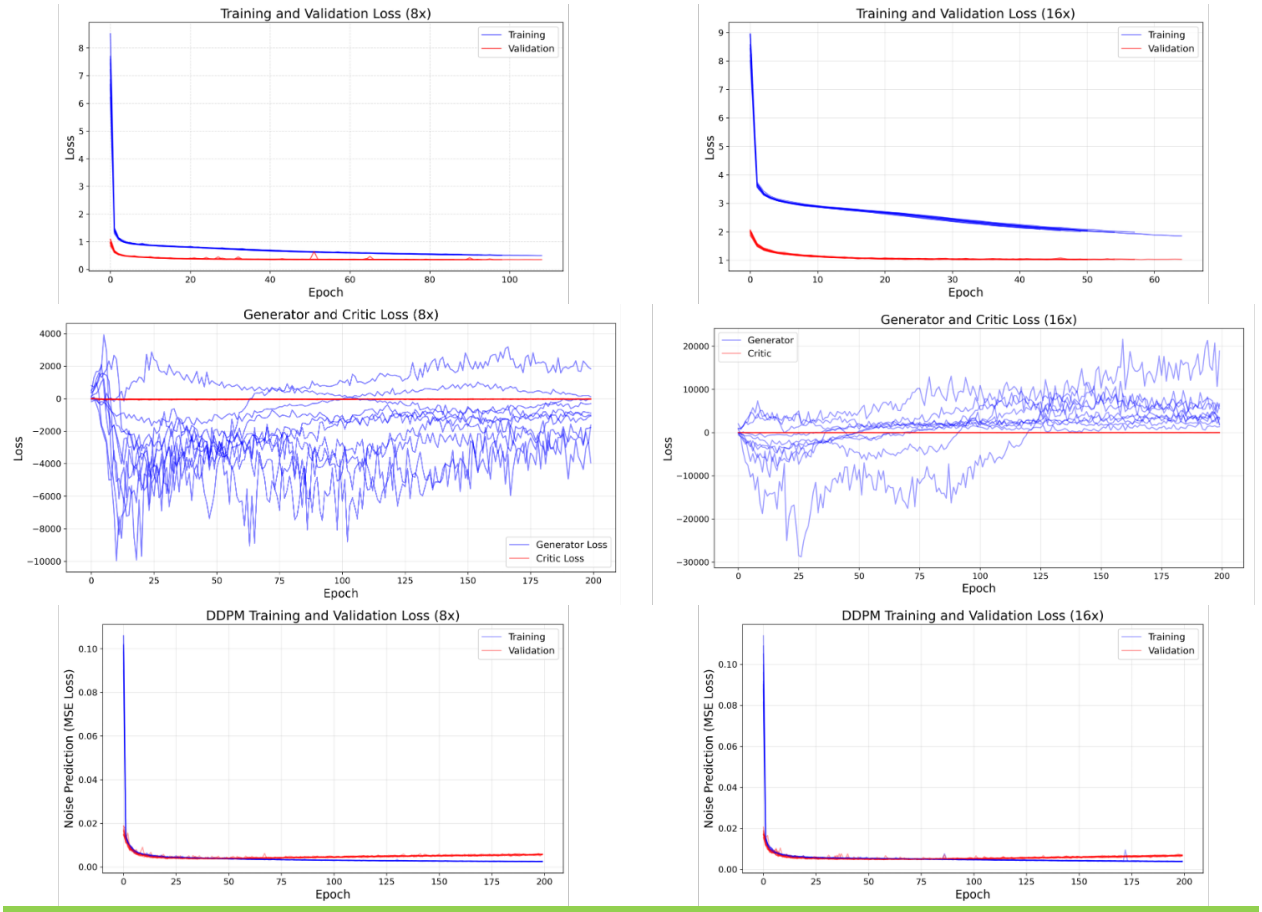

Figure S2. Training and validation curves for the U-Net, WGAN, and DDPM models under 8× and 16× downscaling configurations. (top) Training and validation loss for U-Net; (middle) Generator and critic losses for WGAN; (bottom) Noise-prediction MSE loss for DDPM. Each subplot represents one of ten independent (random) initialization seeds, shown separately for 8× (left column) and 16× (right column). Blue line represents training loss behavior (generator loss in case of WGAN) whereas red line represents validation loss behavior with epochs (critic loss in case of WGAN).

To address these concerns, we have expanded the description of the training, validation, and test data partitioning and revised the discussion associated with Figure S2. For the U-Net, convergence was monitored using validation loss, and the best-performing checkpoint was

selected through early stopping with a patience of 10 epochs based on validation performance. The corresponding training and validation curves exhibit rapid convergence and remain closely aligned throughout training, indicating stable optimization and strong generalization.

For the DDPM, we have revised the discussion of the training behavior to avoid overinterpreting the loss evolution. The training and validation noise-prediction losses remain closely aligned throughout training for both the 8× and 16× configurations, indicating stable convergence and no clear evidence of substantial overfitting. While convergence is slower than for the U-Net due to the iterative denoising process, performance evaluation and model comparison are based on diagnostics computed from an independent test dataset rather than training loss alone.

Regarding the WGAN, we note that adversarial training involves separate generator and critic objectives whose loss values are not directly comparable to conventional supervised validation losses and are often only weakly correlated with sample quality. Consequently, WGAN convergence and model selection were not assessed solely from adversarial losses. Instead, we monitored multiple validation diagnostics, including SSIM, exceedance probability characteristics, spatial correlation statistics, and qualitative sample realism. The critic and generator losses remained bounded throughout training across all seeds, with no evidence of training instability or mode collapse.

Overall, model generalization was evaluated using an independent test dataset and multiple complementary diagnostics, including exceedance probability, spatial autocorrelation, temporal autocorrelation, radial power spectra, and reconstruction statistics. We have revised the manuscript to clarify these procedures and to provide a more balanced interpretation of the training curves.

"Nevertheless, the loss trajectories remain bounded across all seeds, with no evidence of training instability or mode collapse. It is important to note that adversarial loss values are not directly comparable to conventional supervised validation losses and are often only weakly correlated with sample quality. Therefore, WGAN convergence was assessed not only through adversarial losses but also through independent validation diagnostics, including SSIM, exceedance probability characteristics, spatial correlation statistics, and qualitative sample realism. Collectively, these results indicate stable adversarial training across both super-resolution configurations." (see line 458-465)

Comment 2:

Directly related to the previous point: Sample Size vs. Model Complexity (Critical)

The technical rigor of the study is challenged by the potential imbalance between the available data and the model's capacity. The authors utilize ~33,000 daily fields (with only ~19,200 samples for training) to train high-capacity architectures. Modern WGAN and DDPM implementations (including features mentioned like

sinusoidal time embeddings and FiLM layers) often contain millions of trainable parameters.

- The authors must provide a table explicitly summarizing the total number of trainable parameters for the U-NET, WGAN, and DDPM.
- Given the relatively small training set size and the high complexity of the models, the risk of overfitting is severe. The authors must discuss the generalization potential of these models in this context.

Conclusion on the first two points:

This evidence of potential overfitting is a major flaw that propagates through all results and discussions in the manuscript. Before any further analysis can be considered, the authors must demonstrate that the models (especially the DDPM) are not over-parameterized for the provided dataset and that the reported performance metrics are not the result of a model that has failed to generalize.

Response: We thank the reviewer for raising this important concern regarding model complexity relative to training sample size. We agree that reporting model capacity explicitly is necessary for a rigorous comparison. To address this issue, we have added a table summarizing the number of trainable parameters for all models evaluated in this study (Table S1).

Table S1. Number of trainable parameters for the architectures evaluated in this study, including the U-Net baseline, the WGAN generator and critic, and the DDPM denoising U-Net.

Model	Trainable Parameters
U-Net	497,121
WGAN Generator	497,121
WGAN Critic	552,417
DDPM Denoising U-Net	14,567,649

The larger parameter count of the DDPM primarily arises from the conditional denoising U-Net backbone and the additional timestep-conditioning components used during the diffusion process. However, model size alone does not necessarily imply overfitting. Generalization was evaluated using an independent test dataset that was not used during training or model selection. Across all three frameworks, performance was assessed exclusively on unseen test samples using multiple complementary diagnostics, including exceedance probability, radial power spectra, spatial autocorrelation, temporal autocorrelation, and reconstruction statistics. The consistent behavior observed across these independent evaluations suggests that the models learned transferable precipitation features rather than memorizing training examples.

We also note that the downscaling problem considered here is a supervised conditional mapping from coarse-resolution precipitation fields (8×8 or 16×16) to high-resolution targets (128×128). This setting is substantially more constrained than unconstrained image generation tasks, reducing the effective risk of memorization and helping regularize the learning process.

Additional concerns

Comment 3:

Perfect-Model Framework and Practical Usability (Lines 154–157): The authors describe their approach as a "perfect-model super-resolution framework" where both inputs and targets originate from the same dataset (ERA5-Land). While this allows for controlled evaluation, it ignores the critical real-world challenge of "predictor mismatch" or bias between different datasets (e.g., GCM output vs. regional simulations). This design choice significantly hinders the practical usability of the proposed models, as they are not tested against actual biased low-resolution information found in climate model outputs. This limitation must be explicitly stated in the **Abstract**, and discussed in the **Introduction**, and **Conclusion**. The authors should clarify how these models would perform when the "trace" of mesoscale information is truly absent or biased in the low-resolution input.

Response: We thank the reviewer for this thoughtful comment. We agree that the present study adopts a controlled perfect-model super-resolution framework, in which both the coarse-resolution inputs and high-resolution targets are derived from the same ERA5-Land dataset. This setup does not fully represent operational climate downscaling applications, where coarse predictors often originate from external sources such as GCMs or regional climate models and may contain systematic biases, structural errors, or missing mesoscale information.

Our motivation for using this framework was to enable a fair and controlled comparison of the three deep-learning approaches (U-Net, WGAN, and DDPM) under identical input-output conditions. By minimizing predictor mismatch as a confounding factor, the study isolates differences attributable to the learning frameworks themselves and allows a clearer assessment of their ability to reconstruct fine-scale precipitation structure, spatial dependence, extremes, and uncertainty characteristics.

We agree that strong performance in a perfect-model setting does not automatically guarantee equivalent skill when applied to biased climate-model predictors. In practical applications, additional challenges such as bias correction, domain adaptation, predictor selection, and limited mesoscale information must be addressed. Furthermore, when mesoscale information is weak, absent, or poorly represented in the low-resolution predictors, the achievable reconstruction skill will likely decrease regardless of the downscaling architecture employed.

To address this concern, we have expanded the discussion of study limitations and future research directions in the revised manuscript. We now explicitly acknowledge that the present results should be interpreted as a controlled benchmark of model behavior within a perfect-model framework rather than a direct demonstration of readiness for operational climate-model downscaling. We also highlight the extension of these methods to biased climate-model predictors as an important direction for future work.

See revised text between line 856 and 885, "Several limitations should be acknowledged. The models were trained and evaluated using ERA5-Land as the high-resolution reference, which, although widely used, remains a reanalysis product rather than a direct observational dataset. The analysis was also restricted to precipitation, and the relative advantages of generative approaches may not necessarily extend to multivariate downscaling problems involving additional variables. Furthermore, the models did not explicitly incorporate physiographic information such as topography and land-surface characteristics, which may influence performance in regions where terrain strongly affects precipitation patterns. The architectures evaluated here should be viewed as representative implementations rather than fully optimized realizations. In particular, the baseline U-Net employed a conventional MSE objective, whereas precipitation fields exhibit intermittency, skewness, and heavy-tailed behavior that may be better represented by distribution-aware loss functions (e.g., Tweedie, quantile-based, or probabilistic objectives; (Hunt, 2025; Rastogi et al., 2025)). Finally, all experiments were conducted under historical climate conditions, and the robustness of these approaches under future climate states and associated distributional shifts remains an important area for future investigation.

Future work should address these limitations progressively. Evaluating model performance against observation-based targets, incorporating physiographic conditioning variables, and extending the framework to multivariate downscaling problems are all important near-term directions. Further research should also explore the sensitivity of conclusions to architectural design choices, loss functions, data transformations, and conditioning strategies. Assessing robustness under future climate conditions will likely require complementary approaches such as domain adaptation, physically informed constraints, and hybrid frameworks that combine the flexibility of generative deep learning with the dynamical consistency of process-based climate models. Overall, the results demonstrate that generative approaches, particularly WGANs and DDPMs, offer substantial advantages over conventional deterministic downscaling for reproducing precipitation variability, extremes, and uncertainty, while also highlighting the importance of evaluating model performance across multiple complementary diagnostics rather than relying on a single metric."

Comment 4:

Data Normalization: The manuscript mentions a log(1+x) transformation and min-max normalization specifically for the DDPM (Line 258). It is unclear if any normalization or transformation was applied to the precipitation data for the U-NET and WGAN models. Given the heavy-tailed nature of precipitation, this is a critical detail. If no normalization was used

for the other models, the authors must justify this choice or clarify the preprocessing steps for all architectures.

Response: Thank you for highlighting this important point. We agree that preprocessing and normalization choices are particularly relevant for precipitation data because of its highly skewed and heavy-tailed distribution. To clarify, the U-Net and WGAN models were trained directly on precipitation fields in the transformed data space used throughout the study and did not require an additional target normalization step. In contrast, DDPM training involves iterative noise addition and removal over hundreds of denoising steps, making training substantially more sensitive to the numerical scale of the target variable. For this reason, the DDPM experiments employed a log(1+x) transformation followed by min-max normalization prior to diffusion training.

We have revised the manuscript to clarify that the additional normalization step was specific to the DDPM framework. (See line 266-270), "Unlike the U-Net and WGAN models, which were trained without an additional target normalization step, DDPM training employed a log(1+x) transformation followed by min-max normalization. This normalization was introduced to improve numerical stability during the iterative diffusion and denoising process."

Comment 5:

Stochastic vs. Epistemic Uncertainty: The authors use 10 independently trained models to form an ensemble (Line 270), which they state reflects "epistemic uncertainty." However, the primary strength of generative models like WGAN and DDPM is their ability to produce multiple stochastic realizations from a single trained model. The authors should evaluate the generative potential/stochasticity of a single trained model rather than relying solely on an ensemble of differently trained models, as the latter is computationally expensive and less practical for end-users.

Response: We thank the reviewer for this important observation. We agree that uncertainty arising from stochastic sampling of a single trained generative model (aleatoric variability) is conceptually distinct from uncertainty arising from independently trained model initializations (epistemic variability). To address this concern, we have added a new subsection entitled **"5.2.6 Stochastic Variability and Uncertainty"** in the revised manuscript (Lines 692–726). In this analysis, we explicitly compare (i) within-seed variability obtained from repeated stochastic realizations generated by a single trained model and (ii) across-seed variability obtained from independently trained models initialized with different random seeds. The comparison is performed using spatial autocorrelation, exceedance probability, and power spectral density diagnostics. The results show that both sources of variability affect precipitation statistics, but across-seed variability produces a broader and more persistent spread, particularly in the upper tail of the precipitation distribution. These findings clarify the distinct contributions of sampling

variability and training variability and provide a more rigorous interpretation of ensemble uncertainty in generative precipitation downscaling models.

The updated text between lines 689 and 725 in the revised version is,

5.2.6 Stochastic Variability and Uncertainty

The comparison between within-seed and across-seed ensembles shows that both sources of variability affect the simulated precipitation distribution, but across-seed variability is consistently larger. This distinction is most evident in the exceedance probability curves (Figure 9b1–b2). Within-seed realizations remain relatively clustered at moderate thresholds but begin to diverge in the upper tail, particularly beyond approximately 150 mm/day. In contrast, across-seed ensembles show a broader spread even at lower precipitation thresholds, and this spread increases further toward the most extreme events. This indicates that stochastic sampling within a fixed trained model contributes to tail uncertainty, but differences among independently trained model initializations produce a stronger and more persistent variability. The lagged spatial autocorrelation characteristics were highly consistent across independent model initializations (Supplementary Figure S8), indicating that the learned spatial dependence structure is robust to training variability. Although the spread among realizations increased slightly at larger lags, particularly for DDPM, the median correlations remained stable across seeds and closely followed the target spatial autocorrelation decay.

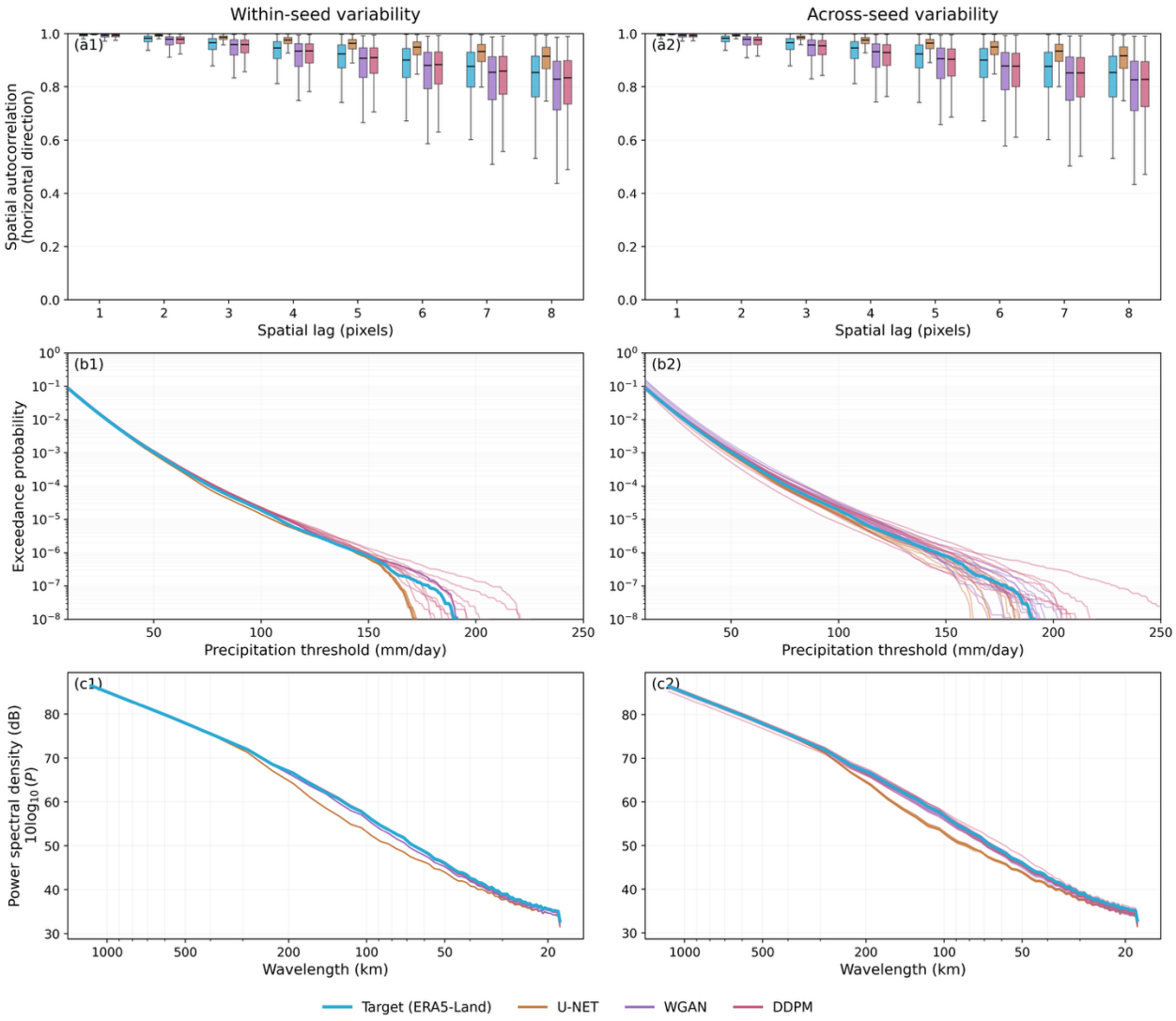

Figure 9: Within-seed and across-seed variability of U-Net, WGAN, and DDPM predictions over the independent Northeast test region at 16× super-resolution. Panels (a1–a2) show boxplots of horizontal spatial autocorrelation at lags 1–8 pixels, where within-seed variability is estimated from multiple stochastic realizations generated by a single trained model and across-seed variability is estimated from independently trained model initializations. Panels (b1–b2) present exceedance probability curves as a function of precipitation threshold (mm/day), illustrating variability in the representation of moderate-to-extreme precipitation intensities. Panels (c1–c2) show radial power spectra describing the distribution of spatial variance across wavelengths. Colored lines and boxplots represent individual model realizations, while the target (ERA5-Land) reference is shown for comparison. Left-column panels correspond to within-seed variability and right-column panels correspond to across-seed variability.

For spatial autocorrelation (Figure 9a1–a2) and radial power spectra (Figure 9c1–c2), the contrast between within-seed and across-seed behavior is less pronounced. Both ensemble configurations show broadly similar spatial dependence and spectral

characteristics, suggesting that the learned spatial organization is relatively stable. Nevertheless, the across-seed boxplots show slightly larger dispersion at longer spatial lags, indicating that model initialization also affects spatial dependence, although less strongly than it affects precipitation extremes. These results suggest that uncertainty in generated precipitation fields is metric dependent. Spatial structure is comparatively robust to both stochastic sampling and independent model initialization, whereas distributional tail behavior is more sensitive, with across-seed variability representing the dominant source of uncertainty."

Comment 6:

Static Variables: Were any static variables (e.g., elevation, land cover, distance to coast) included as auxiliary inputs to the models? Orographic forcing is mentioned as a relevant driver (especially for one of the domains/regimes line 127), but looks like the models were trained without providing them with this information. Having access to the underlying topography can assist in the downscaling process: why did the authors choose not to include any static covariants in the training?

Response: We thank the reviewer for this insightful comment. We agree that physically meaningful conditioning variables, such as temperature, humidity, wind fields, topography, and land-surface characteristics, can provide valuable information for precipitation downscaling and may improve model performance. However, the primary objective of this study was to conduct a controlled intercomparison of three widely used generative AI architectures (U-NET, WGAN, and DDPM) within a perfect-model super-resolution setting. Restricting the input to coarse-resolution precipitation allowed us to isolate differences attributable to model architecture and training strategy while avoiding confounding effects arising from differing predictor selections.

We agree that incorporating physically meaningful conditioning information represents an important step toward more realistic climate downscaling applications. Such information could be incorporated through additional predictor channels or dedicated conditioning mechanisms and may further improve the representation of precipitation occurrence, intensity, and spatial structure. We have therefore expanded the discussion of study limitations and future research directions to acknowledge this point and highlight physics-informed and multivariate conditioning frameworks as promising avenues for future work.

We added the following discussion to the limitations section (Lines 858–863):

"The analysis was also restricted to precipitation, and the relative advantages of generative approaches may not necessarily extend to multivariate downscaling problems involving additional variables. Furthermore, the models did not explicitly incorporate physiographic information such as topography and land-surface characteristics, which may influence performance in regions where terrain strongly affects precipitation patterns."

We also expanded the future directions section (Lines 872–875):

"Future work should address these limitations progressively. Evaluating model performance against observation-based targets, incorporating physiographic conditioning variables, and extending the framework to multivariate downscaling problems are all important near-term directions."

Comment 7:

3. Specific Comments and Technical Corrections

Hydrologic Relevance: The authors mention "hydrologic modeling" as a motivation many times in the text. Could the authors clarify the specific "hydrologic" metrics or assessments used beyond standard meteorological diagnostics?

Response: Thank you for this comment. Beyond standard meteorological diagnostics, our evaluation includes several precipitation metrics with clear hydrologic significance. These include the probability of zero precipitation (wet/dry occurrence), mean and variance, which are relevant for water balance and rainfall variability, as well as L-moments (L-dispersion, L-skewness, and L-kurtosis), which characterize distributional shape and heavy-tailed precipitation behavior.

We also assess exceedance probability curves, which are directly relevant for heavy-rainfall frequency and flood-generating events. In addition, spatial correlation diagnostics are important for understanding storm organization and basin-scale runoff coherence, while temporal autocorrelation helps characterize multi-day persistence, storm sequencing, and antecedent wetness conditions that strongly influence hydrologic response.

Comment 8:

Target Data Selection (Lines 115–122): The authors use ERA5-Land as the target data. Please provide a brief justification for using a reanalysis product as "Ground Truth" (GT) rather than observational-based data (e.g., high-resolution radar products).

Response: Thank you for this important comment. We agree that ERA5-Land is a reanalysis product rather than a direct observational dataset, and therefore should not be interpreted as a perfect representation of true precipitation observations. We selected ERA5-Land as the high-resolution reference because the objective of this study was to conduct a controlled intercomparison of U-Net, WGAN, and DDPM within a perfect-model super-resolution framework. ERA5-Land provides spatially complete, internally consistent, long-term gridded precipitation fields at sufficiently high spatial resolution, making it well suited for generating paired low- and high-resolution samples through the block-averaging procedure used in this work. This framework allows differences among model architectures to be isolated without introducing additional uncertainties associated with observational errors, data gaps, differing spatial supports, or mismatches between observational and model-based products.

We acknowledge, however, that strong performance relative to ERA5-Land does not necessarily imply equivalent performance relative to independent observations. To clarify this limitation, we have added text to the Conclusions noting that ERA5-Land is a reanalysis-based reference dataset and that future work should evaluate these architectures against observation-based products and operational downscaling targets.

Comment 9:

DDPM Hyperparameters (Section 2.2.3): training details for the DDPM are missing. Specifically, please provide the optimizer used (e.g., Adam, AdamW) and the specific learning rate or learning rate schedule employed during training.

Response: Thank you for pointing this out. We have revised the manuscript to explicitly report optimizer hyperparameters for all models. The U-Net uses Adam with default $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and no weight decay. The WGAN already used non-default values ($\beta_1 = 0.0$, $\beta_2 = 0.9$), which are now consistently reported alongside other models. The DDPM section has been updated to specify the use of AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-4}.

The text has been updated to, "The model is trained end-to-end using the Adam optimizer (learning rate 1×10^{-4}, $\text{beta}_1 = 0.9$, $\text{beta}_2 = 0.999$, batch size 32) with early stopping when validation loss fails to improve for ten consecutive epochs (patience = 10). No weight decay is used, and the best checkpoint based on validation set MSE is retained for evaluation." (See line 194-198)

Comment 10:

Figure 3 Clarification: What do the columns in Figure 3 represent? Please label them clearly in the figure or the caption. The low-resolution input in the last column does not appear to match the high-resolution Ground Truth (GT). Please verify if the block-averaging or interpolation used to create the low-resolution inputs is consistent across the visualization.

Response: We thank the reviewer for this comment. Each column in Figure 3 represents an independent precipitation event randomly selected from the test set. The rows correspond to the low-resolution input, the high-resolution ERA5-Land target, and the corresponding predictions from the U-Net, WGAN, and DDPM models. To improve clarity, we have revised the figure caption to explicitly describe the organization of the panels.

Regarding the apparent mismatch between some low-resolution inputs and high-resolution targets, we verified that the low-resolution fields were generated directly from the corresponding high-resolution ERA5-Land targets using the same block-averaging procedure employed during model training. The apparent discrepancies arise because coarse-resolution aggregation substantially smooths and compresses localized precipitation features, particularly for fragmented or intermittent precipitation events. In addition, the low-resolution inputs are displayed on their native coarse grid, whereas the target and

model outputs are shown at the full high-resolution scale, which can further accentuate visual differences.

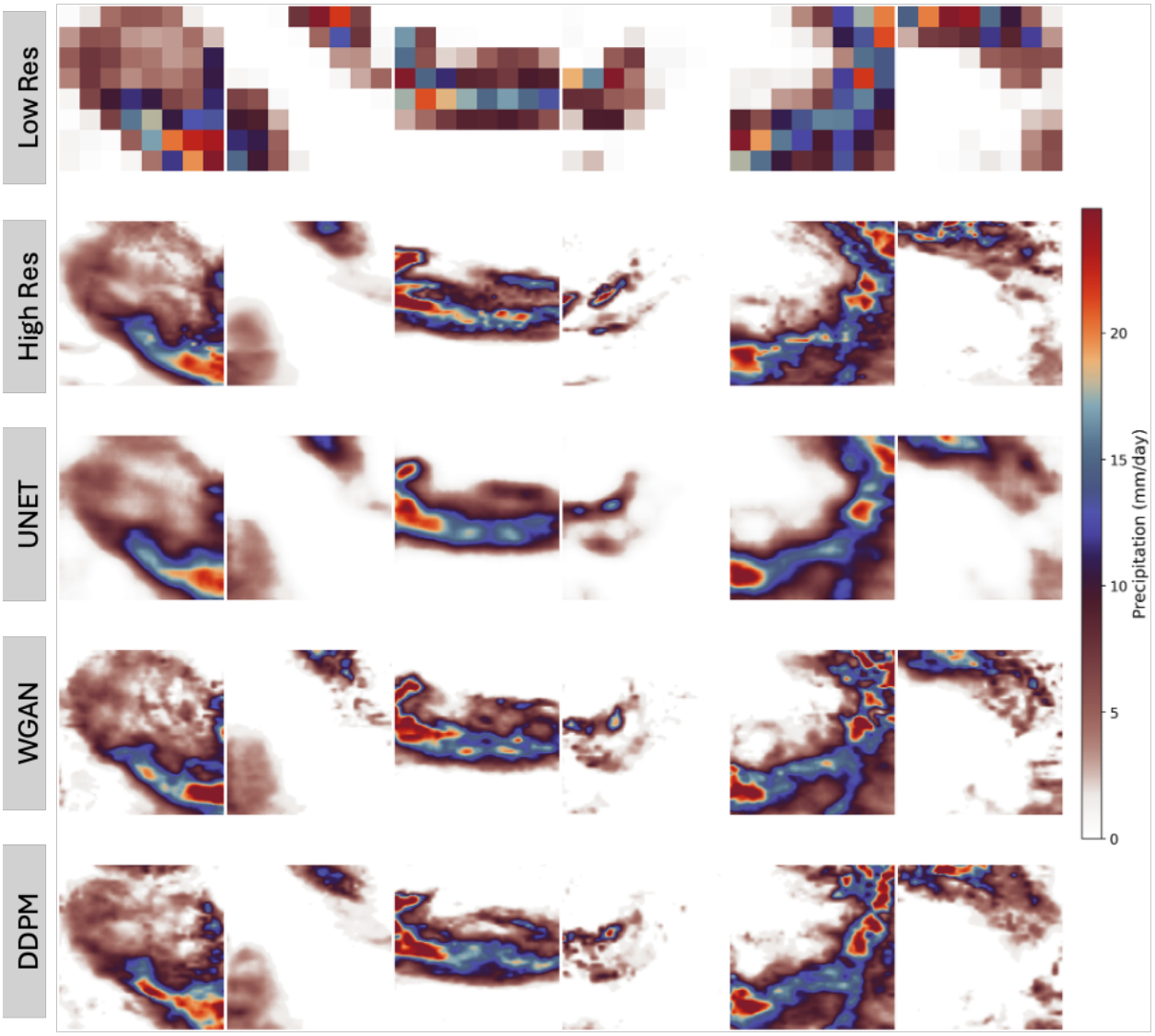

Figure 3. Reconstruction of high-resolution precipitation fields for the 16× downscaling task (8×8 → 128×128). Columns show six randomly selected test samples, while rows show the low-resolution input, high-resolution target (ERA5-land), and predictions from U-NET, WGAN, and DDPM respectively.